Cleaning Messy Data

Online course learning objectives

This course will introduce the fundamentals of cleaning messy data. It will provide a clear understanding about what messy datasets are and why they need to be cleaned, as well as giving lots of practical examples for cleaning data sets in different programs.

This course will help learners to:

- Recognize when data are messy and require cleaning
- Apply cleaning methods to messy datasets
- Understand how cleaning messy data contributes to good data management
- Perform quality control of data

Language: English
Time to complete: 2 hours
Level: Beginner
Instructor: Dr Alessandra Vigilante

Online course full syllabus

MODULE ONE: HELP! MY DATA ARE MESSY

Even the most organized person can make mistakes when recording and saving data. At first, datasets can look clean and reproducible but as soon as we try to add more data or use them for analysis or visualization purposes, issues begin to arise, and we find ourselves needing to clean the data! In this module, you will learn what messy data are, and why it's so important to recognize and clean them as soon as possible (and avoid them in the future!).

This module will help you to:

- Recognize what is meant by "messy data"
- Identify when quantitative and qualitative data are messy
- Predict common errors made while dealing with data

MODULE TWO: WHY CLEAN MESSY DATA?

Messy data will waste your time, will confuse your collaborators, and will certainly negatively impact your analysis and your research output.

Cleaning Messy Data

In this module, we'll explain why it's so important to have clean data you can trust, both to obtain reliable results and for creating sustainable and interoperable datasets.

This module will help you to:

- Recognize how dealing with messy data makes analysis more complex
- Discover the importance of the FAIR principles
- Identify that messy data lead to inaccurate conclusions
- Appreciate cleaning data as a skill for employability

MODULE THREE: HOW CAN I CLEAN MY MESSY DATA?

Most of the time, quantitative data are recorded and saved in text files using a spreadsheet program. Excel isn't the only spreadsheet program, but it's arguably the most used one. Free spreadsheet programs include LibreOffice Calc and Apple Numbers for Apple users.

This module will provide background information on different spreadsheet programs and share key skills that can be used to manually clean messy data.

This module will help you to:

- Develop key skills for manually cleaning data in a spreadsheet
- Recognize and avoid formatting problems
- Master basic Excel terminology and concepts
- Set up quality control systems to keep data clean
- Familiarize yourself with different spreadsheet programs